Proceedings of the Twelfth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024)

Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia

Stefano Marchesin[1], **Gianmaria Silvello**[1], **Omar Alonso**[2*]

[1]Department of Information Engineering, University of Padua, Padua, Italy
[2]Amazon, Palo Alto, California, USA
stefano.marchesin@unipd.it, gianmaria.silvello@unipd.it, omralon@amazon.com

Abstract

Knowledge Graphs (KGs) are essential for applications like search, recommendation, and virtual assistants, where their accuracy directly impacts effectiveness. However, due to their large-scale and ever-evolving nature, it is impractical to manually evaluate all KG contents. We propose a framework that employs sampling, estimation, and active learning to audit KG accuracy in a cost-effective manner. The framework prioritizes KG facts based on their utility to downstream tasks. We applied the framework to DBpedia and gathered annotations from both expert and layman annotators. We also explored the potential of Large Language Models (LLMs) as KG evaluators, showing that while they can perform comparably to low-quality human annotators, they tend to overestimate KG accuracy. As such, LLMs are currently insufficient to replace human crowdworkers in the evaluation process. The results also provide insights into the scalability of methods for auditing KGs.

Introduction

Knowledge Graphs (KGs) form the backbone of various downstream tasks, such as search, recommendation, and question-answering (Reinanda, Meij, and de Rijke 2020; Samadi et al. 2015), or applications for virtual assistants (Ilyas et al. 2023; Mohoney et al. 2023). KGs are also used by Large Language Models (LLMs) as additional sources of external knowledge and contextual information in several settings, including chain-of-thought prompting and retrieval-augmented generation. Recent research is focusing on integrating LLMs and KGs to enhance question answering systems by augmenting the factual knowledge of LLMs (Pan et al. 2023). High-quality KGs positively influence these tasks, while low-quality KGs can diminish their effectiveness. Hence, auditing KG quality and accuracy is crucial given its impact on a wide range of applications.

However, estimating the quality of KGs is challenging due to their large size, making the manual annotation of all KG facts – that is, its triplets – with correctness labels prohibitively expensive. Indeed, real-life KGs like DBpedia (Auer et al. 2007), Wikidata (Vrandecic and Krötzsch 2014), YAGO (Hoffart et al. 2013; Suchanek et al. 2024), and NELL (Mitchell et al. 2018) encompass millions to billions of facts. Besides, we need to consider that the evaluation of a KG is not a one-time event, but rather an ongoing process due to the continuous updates of its facts. For instance, companies are acquired, presidents are elected, and products are updated, thereby requiring regular assessment for quality control.

To overcome this challenge, efficient methods that estimate KG accuracy over a (relatively) small sample are required. To this end, sampling and estimation techniques are central (Cochran 1977), and active learning strategies (Settles 2009) providing quality guarantees emerge as the most promising solution (Gao et al. 2019; Qi et al. 2022; Marchesin and Silvello 2024). These strategies ensure an evaluation process that is both cost-effective and representative of the entire KG – thus offering a practical solution for auditing KG accuracy over time. In this work, we assess the accuracy of a KG by also considering the utility of the facts it contains. Naturally, different parts of the KG can have different degrees of utility depending on the specific downstream task. As an example, let us consider entity-oriented search (Balog 2018), which is the search paradigm of organizing and accessing information centered around entities, their attributes, and relationships. In this task, popular entities in the KG typically have the highest query load (Ilievski, Vossen, and Schlobach 2018; Garigliotti et al. 2019). Therefore, we might want to prioritize their assessment given their significant impact on the search experience compared to less popular entities (Marchesin, Silvello, and Alonso 2024).

Utility can also come in handy when a limited annotation budget is available – a common scenario in real-world applications. Utility can help determining how to allocate annotation resources to maximize the return for the downstream task. Additionally, it can assist in deciding whether to activate filtering and/or correction mechanisms in low-quality situations (Paulheim 2017; Faralli, Lenzi, and Velardi 2023).

Motivated by these observations, we propose an efficient, utility-oriented KG accuracy evaluation framework that can scale to the size of real-life KGs with limited human annotations and strong statistical guarantees. The proposed framework can also be used to prioritize the evaluation dimension we decide to focus on, such as accuracy, quality or utility.

Figure 1 presents the framework and its components. In ❶, a utility model is defined based on the KG and web

*Work does not relate to the author's position at Amazon.

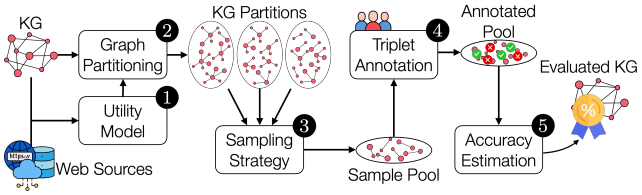

Figure 1: Utility-oriented KG accuracy evaluation.

sources. The model is used to annotate KG triplets with scores that reflect their utility for downstream applications. In ❷, the KG undergoes a partitioning procedure that divides KG triplets into separate subsets according to their utility scores. In ❸, batches of samples are collected from each partition according to a given sampling strategy. In ❹, sampled triplets are routed to multiple annotators to gather correctness labels. In ❺, given the accumulated annotations, an estimator computes an unbiased estimate of the KG accuracy and the corresponding Confidence Interval (CI).

The main **contribution** of this work is the design and implementation of the proposed framework in a real-world scenario. We operationally test the framework on DBpedia and introduce a method to partition DBpedia based on facts utility scores. We set up a crowdsourced annotation task involving three expert annotators and 60 layman annotators, providing, to the best of our knowledge, the first extensive human-based accuracy evaluation of DBpedia. Subsequently, we use these human-provided annotations to test the proposed evaluation framework. Additionally, we instantiate the evaluation framework by employing three open-source LLMs instead of human assessors to evaluate the extent to which LLMs can replace human annotators for KG accuracy evaluation at scale.

The main **outcomes** of this work are:

- An analysis of human annotator performance on KG accuracy evaluation, highlighting the effectiveness of the considered crowdsourcing strategies.
- A two-level estimation of DBpedia accuracy, providing estimates with strong statistical guarantees for both utility-derived partitions and the entire KG.
- An assessment of LLMs as KG evaluators, revealing that while LLMs perform comparably to low-quality human annotators, they tend to overestimate accuracy scores and are therefore not yet suitable substitutes for human crowdworkers in auditing KG accuracy.
- We release all the collected data in anonymized format.[1]

Preliminaries

Notation. A KG is a directed, edge-labeled multi-graph, usually defined as $G = (V, R, \eta)$, where $V = \{E \cup A \cup B\}$ is the set of nodes in G, with E as entities, A as attributes, and B as blank nodes; R is the set of relationships between nodes in G; and $\eta : R \to (E \cup B) \times (E \cup A \cup B)$ is a function assigning an ordered pair of nodes to each relationship. The η function produces the ternary relation T of G (Bonifati et al. 2018). Without loss of generality, we consider ground RDF graphs – that is, RDF graphs without blank nodes. As a result, we can redefine η as $\eta : R \to E \times (E \cup A)$. Thus, the ternary relation T becomes the set of (s, p, o) triplets such that $s \in E$, $p \in R$, and $o \in E \cup A$, where $M = |T|$ is its size. In this work, we consider triplets as first-class citizens along with nodes and relationships. Therefore, we can redefine a KG as $G = (V, R, T, \eta)$, and define an entity cluster $G[e] = \{(s, p, o) \in T \mid s = e\}$ as a set of triplets in $T \in G$ sharing the same subject $e \in V$. Triplets whose object is an entity are called triplets with entity property, whereas those with attribute objects are known as triplets with data property. We also refer to a triplet as a fact.

KG. We consider the 2015-10 English version of DBpedia as G, with 6.2M entities and 1.1B triplets. DBpedia 2015-10 is a popular version that has been used for various entity-oriented downstream tasks (Hasibi et al. 2017; Hasibi, Balog, and Bratsberg 2017; Paranjpe, Bhowmik, and de Melo 2020; Arabzadeh, Bigdeli, and Bagheri 2024).[2] Like Hasibi et al. (2017), we require subject entities to be resources that include both the `rdfs:label` and `rdfs:comment` predicates. Furthermore, we exclude T-Box triplets and only focus on A-Box ones. T-Box encompasses the ontological entities and relationships, whereas A-Box contains the assertions that need to be evaluated for accuracy. After filtering, G consists of 4.6M entities and 170M triplets.

KG accuracy. We define accuracy based on the semantic validity of triplets (Batini et al. 2009), assessing whether the statements they express are correct. Since an atomic fact is either correct or incorrect, we use a binary validation approach for triplets (Esteves et al. 2018), treating all incorrect facts equally regardless of the error type. Then, KG accuracy can be defined as the mean accuracy of its triplets $\mu(G) = \frac{\sum_{t \in T} \mathbb{1}(t)}{M}$, where $\mathbb{1}(t)$ is an indicator function with 1 indicating correctness and 0 incorrectness.

❶ Utility Model

Utility is task-specific and can be tailored to fit particular requirements. For example, Zheng et al. (2022) measure fact utility by their web popularity. In contrast, we propose a utility model based on SPARQL query logs, determining fact utility through their query frequency.

A fact is used in a SPARQL query either if it appears in the result set or contributes to computing the result. To identify the facts in the result set, we simply execute the queries against the KG. For facts used in computation, we calculate query provenance. We adopt the lineage method for data provenance, as described in Dosso, Davidson, and Silvello (2022), because it is widely used, intuitive, and computationally efficient. Lineage is defined as the set of input KG facts that contribute to generating an output fact. To implement this, we transform SPARQL queries into `CONSTRUCT` queries, which retrieve all facts involved in the query responses. By computing the frequency of these facts, we derive a utility score. This score serves as a proxy for user en-

[1] https://github.com/KGAccuracyEval/dbpedia-accuracy-estimation

[2] Many large-scale KGs exist, but most are proprietary. This makes DBpedia one of the most viable open-source alternatives.

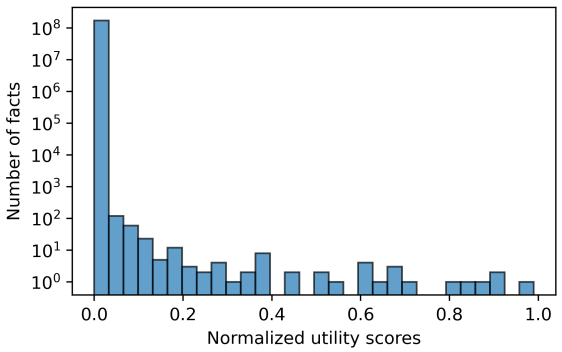

Figure 2: Distribution of facts across utility scores. The scores are normalized to improve presentation clarity.

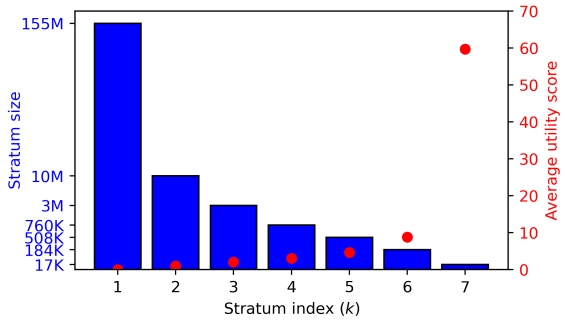

Figure 3: Size and average utility score of the CSRF strata.

gagement, reflecting the popularity and relevance of facts within the analyzed query logs.

As query logs, we exploit those from the Linked SPARQL Queries (LSQ) 2.0 dataset (Stadler et al. 2024), comprising 11.56M unique SPARQL queries extracted from the logs of 27 distinct endpoints. For DBpedia 2015-10, LSQ 2.0 contains 1.7M valid queries that yield non-empty results, primarily consisting of 1.67M `SELECT` queries (98%), alongside a small portion of 31K `CONSTRUCT`, 25 `ASK`, and 6 `DESCRIBE` queries (2%). Figure 2 shows the distribution of facts across normalized utility scores. Most facts cluster near zero, but there is a diverse spread with a notable long tail. A similar trend was observed by Zheng et al. (2022) when using web popularity to compute utility.

❷ Graph Partitioning

We employ stratification to partition the KG into subsets of triplets. Stratification is a statistical technique that divides the population into k subsets, or strata, based on features of interest, ensuring these features are well represented within each subset of the population. When strata are internally homogeneous, meaning minimal variation in measurements across units, precise estimates can be derived from small samples within strata (Cochran 1977).

We perform stratification based on utility scores. As partitioning strategy, we use the Cumulative Square Root of Frequency (CSRF) method (Dalenius and Hodges 1959). CSRF aims to achieve minimal intra-stratum variance in scores and has been used in similar settings given its strong theoretical foundation (Gao et al. 2019; Marchant and Rubinstein 2017, 2021). Once we obtain fact utility scores, we input them into CSRF to derive the partition family $\mathcal{P} = \{P_1, \ldots, P_k\}$.

Groundwork suggests partitioning the KG into a small number of strata $k \in \{2, \ldots, 10\}$ to keep partition sizes considerably large (Gao et al. 2019; Qi et al. 2022). In fact, a large number of strata could result in small sample sizes within each stratum, potentially preventing meaningful conclusions from being drawn and inflating annotation costs.

We set the total number of partitions to $k = 7$ as we found it to be a good compromise between capturing diverse aspects and keeping a reasonable granularity. Figure 3 shows the size and average utility score of the CSRF strata for DBpedia 2015-10. Similarly to the utility distribution (Figure 2), we observe a large stratum with near-zero utility scores and several smaller strata with higher utility scores.

❸ Sampling Strategy

The sampling strategy can significantly impact the efficiency of annotators, influencing the overall cost. In this regard, Gao et al. (2019) show that auditing triplets centered on the same subject entity e – i.e., coming from the entity cluster $G[e]$ – is more cost-effective than auditing triplets involving different (subject) entities. That is, using a Simple Random Sampling (SRS) strategy to collect triplets demands more time from annotators compared to providing them with subsets of triplets focused on the same subject entity.

Thus, we consider Two-stage Weighted Cluster Sampling (TWCS), a state-of-the-art sampling strategy for KG accuracy estimation consisting of two stages (Gao et al. 2019).
Stage 1: sample n entity clusters via Weighted Cluster Sampling, drawing clusters with probabilities $\pi_j = M_j/M$, where $M_j = |G[e_j]|$ is the cardinality of the jth cluster.
Stage 2: sample $\min\{M_j, m\}$ triplets from each jth sampled cluster via SRS without replacement.

To confirm the effectiveness of TWCS over SRS, we ask two experts to conduct two annotation tasks over DBpedia 2015-10, tracking the cumulative time spent for completion. In the first task, we use SRS to draw 50 triplets with distinct subject entities. In the second task, we employ TWCS to first draw entity clusters and then collect (at most) $m = 5$ triplets from each, still totaling 50 triplets. Figure 4 shows the cumulative evaluation time of TWCS versus SRS, confirming TWCS as the most cost-effective solution.

Hence, we use TWCS with second stage size $m = 5$ to collect 500 batches of triplets from each partition P_i of DBpedia, obtaining a sample of 11.62K triplets.

❹ Triplet Annotation

Recruitment process. We chose to build our own team of annotators rather than relying on external crowdsourcing platforms. This decision was driven by the complexity of auditing KG accuracy, which requires a controlled and interactive approach. By recruiting known workers, we ensured continuous interactions and feedback exchanges, cru-

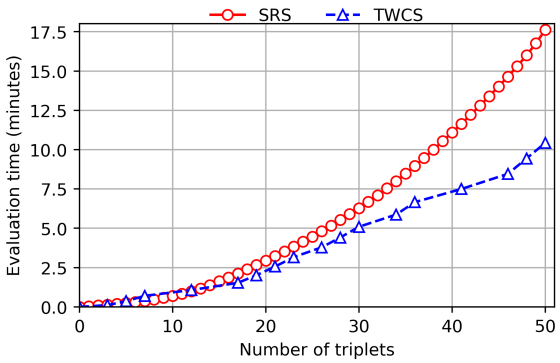

Figure 4: Evaluation cost comparison between SRS and TWCS on DBpedia 2015-10. For SRS, each evaluated triplet is denoted by a circle. For TWCS, each evaluated batch of triplets by a triangle.

cial for enhancing annotation quality. Throughout the annotation period, we provided ongoing training and examples to the annotators. These examples were designed to improve their understanding and skills without influencing their specific decisions.

Our recruitment efforts resulted in a diverse team of 64 layman annotators and 3 expert annotators. The layman annotators were master's students enrolled in a Computational Thinking course, representing a wide range of backgrounds, nationalities, and genders. The expert annotators were computer scientists with experience in crowdsourcing, coming from both academia and industry. This combination of diverse perspectives and expertise enabled us to maintain high standards in the annotation process.

Annotation task. We define the annotation task as labeling the correctness of a batch of triplets sharing the same subject entity. Annotators can choose from three options for each triplet: `Correct`, `Incorrect`, or `I Don't Know` (`IDK`). The `IDK` option helps prevent random judgments. Figure 5(A) illustrates an annotation task where a batch of five triplets about the movie "Jupiter Ascending" contains four correct and one incorrect triplet. Since the considered KG is DBpedia 2015-10, annotators are asked to assume the year is 2015, as a fact that was correct in 2015 may no longer be correct in 2024. Annotators are also instructed to avoid consulting DBpedia or Wikipedia to verify facts, as this would result in evaluating the resource by itself. For incorrect triplets, annotators must specify which element(s) of the triplet (subject, predicate, object) are incorrect. Figure 5(B) displays the error identification task for the incorrect triplet in the "Jupiter Ascending" batch.

Batch routing. Matching triplets with suitable annotators is desirable to improve annotation quality (Zheng et al. 2022). We devise a routing strategy assigning batches of triplets about specific topics to annotators with relevant expertise. To identify the topics we resort to DBpedia categories, specified by the `dcterms:subject` predicate. Since there are more than 20M categories in DBpedia, we re-

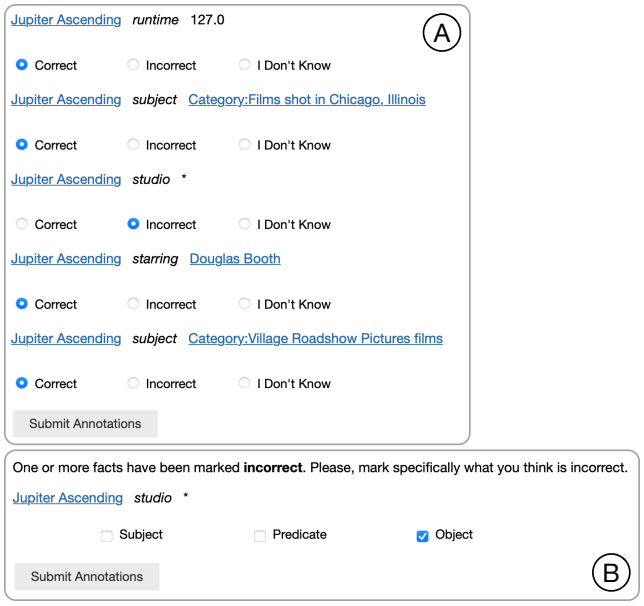

Figure 5: (A) Annotation task for a batch of five triplets about the movie "Jupiter Ascending". (B) Error identification task for the incorrect triplet from the same batch.

strict them to those associated with the subject entities in the sample. On these categories, we employ BERTopic (Grootendorst 2022) to generate interpretable clusters, obtaining 64 distinct clusters, which are manually scrutinized and aggregated to form a set of 11 broad topics. These topics encompass "Architecture", "Business", "Education", "Entertainment", "Geography", "High Tech", "History", "Law and Politics", "News", "Religion", "Science", "Sports", and "Transportation". Annotators are then prompted to indicate which topics they are experts on, with the option to select multiple topics, but at least one. "Entertainment" (56 preferences), "News" (32), and "History" (29) emerge as the most favored topics, while "Religion" (7), "Science" (7), and "Transportation" (3) garner the least popularity.

After defining the topics, we label sampled batches with the topics associated with the corresponding subject entities. Guided by annotators' preferences, we distribute these batches in a round-robin fashion, ensuring even allocation across annotators' areas of expertise. Each annotator is assigned a minimum of 700 triplets, and every triplet is assigned to at least three annotators. This method ensures a fair distribution of tasks and adequate expertise for each triplet.

Quality control. Since the layman annotators are master's degree students, we integrate the annotation tasks into the Computational Thinking course. The students' annotation efforts are rewarded with extra points for their final exam. Each student receives a set of batches to annotate over a six-week period, with the freedom to annotate at their convenience. We track the time and date of each annotation task. For every 100 triplets labeled as `Correct` or `Incorrect`, students earn an extra point – up to a maximum of 5 points. Thus, annotating 500 or more triplets with correctness labels

would yield the full bonus. Points are not awarded for IDK responses, as they cannot be used to compute KG accuracy. However, this may discourage students from using the IDK option in order to maximize points while minimizing efforts.

To ensure high-quality work, we use expert annotations as "honey pots" hidden among students' tasks (Alonso 2019). Two expert annotators independently labeled 100 batches (500 triplets), with a third resolving any disagreements. These expert-annotated batches are split into packets of five and mixed into each student's set. Student annotations on these honey pots are compared with the expert ones. If the label overlap is 60% or greater, the student earns an extra point; otherwise, they do not. Students know honey pots are included but not their exact location, encouraging accurate and unbiased annotations. Unlike crowdsourcing platforms that emphasize speed, our approach uses academic incentives, giving students six weeks for annotations and rewarding high-quality work with extra points for the exam.

We place honey pots every 140 of the (minimum) 700 triplets assigned to each student. This ensures we can obtain reliable estimates of KG accuracy, as previous studies show these can be achieved with annotated samples of circa 500 correctness labels, regardless of the KG size (Gao et al. 2019). By increasing the sample size by 200, we can thus effectively manage cases where students cannot provide correctness labels for all the assigned triplets, while still ensuring they have the opportunity to earn a full bonus.

Label aggregation. There are many aggregation methods available (Hung et al. 2013; Zhang, Wu, and Sheng 2016), with majority voting being one of the most common and efficient in practice (Alonso 2019). Since we have access to annotators' performance on honey pots, we enhance majority voting by implementing a reliability-weighted version.

For each annotator $a \in A$, let Ω_a be the ground truth from the honey pots assigned to them and κ_a the agreement between a and Ω_a, measured by the weighted Cohen's κ coefficient (Cohen 1960). Then, the reliability score for a is $w_a = \frac{1}{1+e^{-\rho\kappa_a}}$, where ρ adjusts the slope of the logistic curve.

We set $\rho = 5$ to ensure a smooth steepness within the $[-1, 1]$ range of Cohen's κ values. This logistic function smooths the κ scores and approximately normalizes them in $[0, 1]$. Thus, a κ score of 0 – indicating random agreement – results in a reliability score of 0.5, thereby reflecting the annotator's random behavior in providing the correct answer.

With these scores, we can define the reliability-weighted majority voting. Let t be a triplet annotated by A_t annotators, and $y_{a,t}$ the label by annotator a. The counters for each label l of triplet t are defined as $C_l(t) = \sum_{a=1}^{A_t} w_a \cdot \mathbb{1}(y_{a,t} = l)$, where $\mathbb{1}(y_{a,t} = l)$ returns 1 if the condition is true and 0 otherwise. The aggregated label L_t for triplet t is the $\arg\max_l C_l(t)$ – that is, L_t is the label with the highest weighted count. In case of ties, we assign $L_t =$ IDK.

❺ Accuracy Estimation

Once triplet annotations have been gathered and aggregated to derive labels, we discard all triplets labeled IDK and feed the rest to an estimator $\hat{\mu}$ to gauge the KG accuracy. Since TWCS is used independently on every partition, we first estimate partition accuracies and then combine them to obtain the KG accuracy. To evaluate KG accuracy, we require $\hat{\mu}$ to be unbiased – i.e., $E[\hat{\mu}] = \mu(G)$. Besides, to quantify the uncertainties inherent in the sampling process, a CI should be provided together with the single-valued point estimate $\hat{\mu}$. To this end, when the sample size is sufficiently large, by the central limit theorem (Casella and Berger 2002), we can construct a $1 - \alpha$ CI as $\hat{\mu} \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\mu})}$, where $z_{\alpha/2}$ denotes the normal critical value with right-tail probability $\alpha/2$. We set $\alpha = 0.05$, thereby building 95% CIs.

Partition estimation. For each partition P_i, let us consider the sample obtained via TWCS, comprising $\sum_{j=1}^{n_i} \min\{M_j, m\}$ triplets from n_i entity clusters. By computing the estimated accuracy $\hat{\mu}_{ij}$ of the jth sampled cluster as the mean accuracy of its sampled triplets, we can define the estimator of $\mu(P_i)$ as $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mu}_{ij}$, which is known to be unbiased (Cochran 1977). Given $\hat{\mu}_i$, we can compute the estimation variance as $\mathrm{Var}(\hat{\mu}_i) = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}(\hat{\mu}_{ij} - \hat{\mu}_i)^2$.

KG estimation. We combine the partition accuracies to derive the KG accuracy estimate. Since the KG is partitioned into k non-overlapping strata, and samples are collected independently within each stratum via TWCS, our scenario fits stratified sampling.

Let E_i be the set of N_i entities in the ith stratum, $\mathcal{C}_i = \{G[e] \mid e \in E_i\}$ the ith stratum cluster family, and $M_i = \sum_{j=1}^{N_i} M_j$ its cardinality. By denoting $W_i = M_i/M$ as the ith stratum weight, we can define the estimator of $\mu(G)$ as $\hat{\mu} = \sum_{i=1}^{k} W_i\,\hat{\mu}_i$, which is known to be unbiased (Cochran 1977). Given $\hat{\mu}$, we derive the corresponding estimation variance $\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{k} W_i^2\,\mathrm{Var}(\hat{\mu}_i)$.

Experiments

Annotation statistics. Out of 64 recruited students, 60 provided annotations, totaling 37,546 across 11,296 distinct facts. On average, each student made 626 ± 81 annotations, with a range of 318 to 742. Each student used the IDK label for about 64 ± 58 triplets, or 10% of their annotations. This indicates that the use of honey pots as a deterrence strategy was effective, as students did not avoid using the IDK label.

For the honey pot annotations, the two expert annotators agreed on 77% of the triplets, with a Cohen's κ score of 0.51, indicating moderate agreement. The third annotator resolved 82% of the disagreements, with unresolved ties labeled as IDK. This moderate agreement among expert annotators underscores the challenge of evaluating KG accuracy.

Partition statistics. The majority of IDK triplets were found in partitions with the lowest utility scores. Notably, partitions P_1 and P_2, presenting the lowest scores, account for more than 50% of all IDK annotations. Since utility scores reflect user engagement from query log frequencies, the abundance of IDK annotations in low-utility partitions suggests that less frequently used triplets are more challenging to assess for correctness. This further underscores the

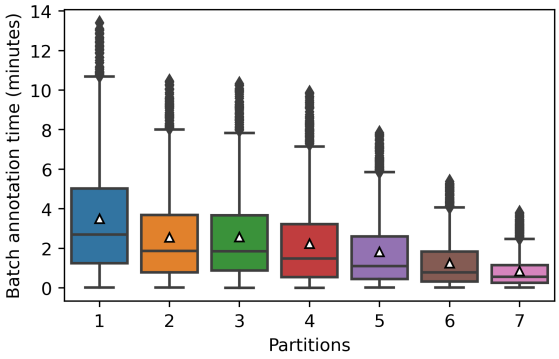

Figure 6: Students' batch annotation time per partition.

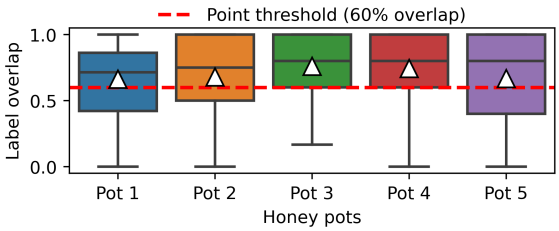

Figure 7: Label overlap between student and expert annotations over honey pots. The dashed red line represents the 60% threshold to grant points to students.

difficulty in auditing the accuracy of large-scale KGs, where most triplets have low utility scores (cfr. Figure 2).

Annotating low-utility triplets proves more challenging and costly. As shown in Figure 6, annotating batches from low-utility partitions demands more time than high-utility ones. Notably, the lowest-utility partition, P_1, requires over double the time of the highest-utility partition, P_7. Given the increased costs, the utility-oriented framework gains significance, enabling strategic resource allocation. Prioritizing annotation efforts on partitions maximizing utility allows focusing on high-traffic or critical areas of the KG, ensuring optimal resource utilization even with limited budgets.

Error statistics. $6,797$ of the $37,546$ collected annotations are about incorrect triplets with the following distribution: $4,205$ annotations (62%) reported errors due to the object, $1,574$ (23%) due to the predicate, 809 (12%) due to both the predicate and object, 157 (2%) due to the subject, and 52 annotations (1%) due to other combinations.

We see that 97% of the reported incorrect annotations involve errors in the object and/or predicate. This is consistent with existing error detection research, which primarily focuses on finding errors in these elements (Paulheim 2017). However, erroneous subjects can still occur.

Quality control data analysis. We used honey pots to evaluate students' annotation quality, offering extra points for their final exam based on performance. Figure 7 displays the overlap between student and expert annotations across their assigned honey pots. Since honey pots were randomly

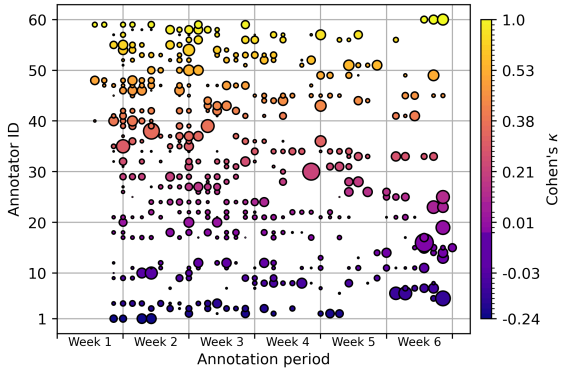

Figure 8: Distribution of students' annotation tasks over the six-week period. Each circle represents a day on which a student conducted annotations; the circle size reflects the number of tasks completed that day.

distributed, sets varied, leading to some partial overlap. Students' overlap with expert annotations exceeded the 60% threshold on average for all honey pots, indicating overall good performance. Notably, the third and fourth honey pots showed the highest performance, while the first and last ones had the lowest. The variation may stem from factors like the early encounter with the first honey pot, causing a "cold start" effect, and the last honey pot appearing towards the end, possibly when students were fatigued or distracted.

Overall, the high overlap scores across all honey pots confirmed that granting extra points for the exam was an effective incentive for high-quality work. The good quality of annotations can also be attributed to students working within their areas of expertise and having a six-week period to complete tasks at their convenience. At the same time, the improvement in annotation quality from one honey pot to the next can be linked to ongoing interactions and feedback provided throughout the annotation period, which enhanced students' understanding and skills.

In this regard, Figure 8 illustrates how students distributed their annotation tasks over the provided time period. Students are sorted from bottom to top by increasing Cohen's κ score, computed over their assigned honey pots. Most students paced their work throughout the weeks, with only a few exceptions completing all their tasks either very early or very late. Most of the students who completed all their tasks near the end had low Cohen's κ scores. This is unsurprising, as they had to conduct an intensive and challenging task close to the given deadline. Hence, they were more likely to become distracted and fatigued compared to students who planned their annotation tasks throughout the six weeks.

KG accuracy estimation. Using student annotations and reliability weights, we aggregated labels for each triplet in the sample. To ensure robustness and quality, only triplets with at least three annotations or two consistent ones were retained, resulting in a dataset of $9,930$ triplets. We excluded triplets labeled `IDK` and used the remaining labels to estimate partition accuracies. These accuracies were then com-

	Correct	Incorrect	IDK	Total	Estimate
P_1	1,851	376	200	2,427	0.83±0.02
P_2	1,416	281	110	1,807	0.85±0.02
P_3	1,343	246	105	1,694	0.85±0.02
P_4	1,164	164	60	1,388	0.89±0.02
P_5	927	133	74	1,134	0.89±0.02
P_6	715	99	22	836	0.90±0.03
P_7	533	96	15	644	0.87±0.03
KG	7,949	1,395	586	9,930	0.83±0.02

Table 1: Distribution of aggregated labels across partitions. For every partition, we report accuracy estimates with CI.

bined to estimate KG accuracy. Table 1 presents the distribution of aggregated labels across partitions.

The distribution of aggregated labels, predominantly `Correct`, indicates that DBpedia 2015-10 is of high quality. Partition accuracy estimates range from a minimum of 0.83, for P_1, to a maximum of 0.90, for P_6. The KG accuracy derived is 0.83, due to the high impact that P_1 size has on the computation of the KG accuracy, being the largest partition. All point estimates have very small CIs, with margins of error never exceeding 3%, thereby providing strong statistical guarantees for the estimation process.

LLMs as KG evaluators. We explored the use of LLMs as KG evaluators to investigate if these tools can automate the KG evaluation process efficiently and reliably. We selected three popular LLMs: Gemma 7B (Mesnard et al. 2024), Llama3 8B (Touvron et al. 2023), and Mistral 7B (Jiang et al. 2023). We chose small models for efficiency, since larger models take longer to produce outputs, making them impractical for evaluating real-life KGs. Also, larger models demand more memory and consume more resources. Nevertheless, beyond our preliminary attempt, there are many other models/configurations to explore.

We used the latest instruction-tuned versions of each LLM, providing them with a prompt containing the same instructions given to students. Specifically, we used few-shot prompting with three examples, one for each possible label, unrelated to sample triplets. Few-shot prompting better approximates the factuality of LLMs and reduces hallucinations in KG evaluation compared to zero-shot or in-domain prompts (Sun et al. 2023). The main difference between the LLM and human annotation tasks was the number of triplets provided. For LLMs, we presented one triplet at a time to simplify the task and increase the precision and quality of the responses. Following Sun et al. (2023), we set the models temperature to zero to make them deterministic and used them without any access to external information sources.

We applied the selected LLMs to the sampled triplets, giving them up to three extra attempts in case they were unable to respond with one of the three required labels. If an LLM failed to provide a proper response for a triplet after four attempts, we set the corresponding label to `IDK`. Then, to determine if LLMs can replace human crowdworkers for KG accuracy evaluation, we compared the Cohen's κ agreement of students and LLMs with experts over the students honey pots. The results of this comparison are shown in Figure 9.

	Gemma 7B	Llama3 8B	Mistral 7B
Compliance	1.00	0.89	1.00
Truthfulness	0.78	0.82	0.90
Informativeness	1.00	0.82	0.66
Accuracy	0.78	0.65	0.59
Balanced accuracy	0.36	0.35	0.46

Table 2: LLM performance over expert annotations.

LLMs outperform 53% of the students on honey pots evaluation, particularly excelling with low-quality students, defined by κ scores below 0.21 (fair agreement). However, not all LLMs consistently surpass students. Mistral outperforms 17 out of 60 students, Gemma 13, and Llama3 only one. Gemma exhibits five outlier cases with perfect agreement with experts on honey pots. These cases involve only `Correct` annotations, making κ inapplicable. Thus, we decided to set the agreement to 1.0. However, this perfect agreement does not reflect Gemma overall quality, as it generally achieves low κ scores. When students attain κ scores above 0.21, LLMs typically maintain low κ scores.

Given the varying quality of the considered LLMs, and since we cannot determine a priori their reliability, a practical solution is to aggregate their outputs through majority vote. This aggregated LLM outperforms 45% of the students, thus providing a potential 45% cost reduction in estimating DBpedia accuracy. Furthermore, the average time required to an LLM to annotate the entire set of sampled triplets is 3.48 hours, compared to 5.21 days for a student. These results suggest that incorporating LLMs into KG accuracy evaluation would yield significant resource savings.

Although the comparison with students on honey pots showed promising results for using LLMs in KG accuracy evaluation, a more comprehensive analysis over the entire set of expert annotations revealed limitations and challenges. For this, we used three LLM-oriented metrics: truthfulness, informativeness, and compliance. Truthfulness and informativeness (Lin, Hilton, and Evans 2022) measure the LLM ability to provide honest (correct answer or `IDK`) and informative (everything but `IDK`) responses. Inspired by Zhou et al. (2023), we defined compliance as the LLM ability to follow instructions, computed as the proportion of times the LLM returns one of the three labels without extra attempts. We also considered accuracy and balanced accuracy.

The evaluation of LLM performance over expert annotations, as shown in Table 2, highlighted three key points.
Compliance: all LLMs adhered to the instructions in the prompt. Both Gemma and Mistral provided proper answers consistently without needing extra attempts.
Truthfulness and informativeness: Llama3 achieved the best balance between truthfulness and informativeness. Gemma had a perfect score for informativeness, never responding with `IDK`, but this overconfidence reduced its truthfulness. In contrast, Mistral was underconfident, resulting in high truthfulness but low informativeness.
Accuracy metrics: although overconfident, Gemma was the

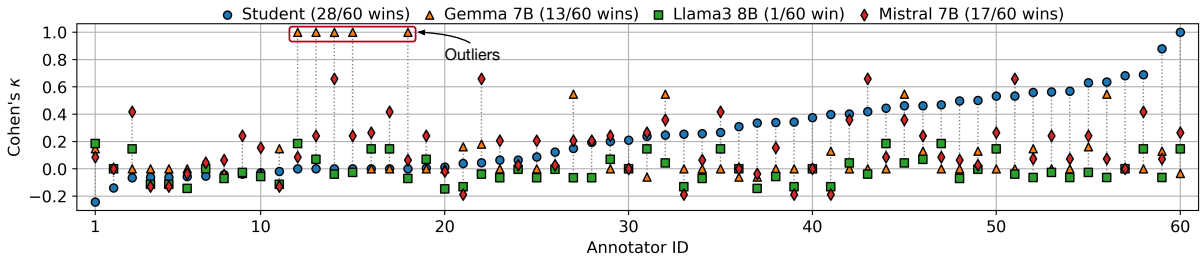

Figure 9: Comparison between LLM and student κ agreements over student honey pots. The outliers in the plot are cases involving only `Correct` annotations from both experts and Gemma, making κ inapplicable. In these cases, we set $\kappa = 1$.

most accurate in terms of raw accuracy. However, this was influenced by the high quality of DBpedia 2015-10 (83%), whose labels are biased towards correctness. This explains why Gemma achieved top accuracy, despite answering with `Correct` in 98% of the cases. Instead, balanced accuracy, which accounts for this bias, revealed Mistral as the most accurate LLM – though all three models had low balanced accuracy, never exceeding 0.46. This low balanced accuracy is due to LLMs reluctance to provide `Incorrect` responses, which made up 18% of the expert annotations. Gemma labeled 2% of the triplets as `Incorrect`, while Llama3 and Mistral labeled less than 1%. Therefore, the LLMs are either biased towards correctness or prefer to reply with `IDK` rather than giving an `Incorrect` response.

This tendency to avoid `Incorrect` responses might be linked to LLMs pretraining data, which likely included most of the information and data stored within DBpedia 2015-10. This lack of `Incorrect` responses led to skewed accuracy estimates when combining the models via reliability-weighted majority voting, resulting in overestimated accuracy scores approaching 100%. Specifically, both partition and KG accuracy estimates reached 0.99, indicating that LLMs are still far from being able to replace human crowdworkers for KG accuracy evaluation.

Related Work

KG accuracy evaluation. KG accuracy evaluation has received limited attention in prior research. Manual evaluation remains the default choice, but the scale of real-life KGs makes it impractical (Xue and Zou 2023). Early efficient approaches either lacked statistical guarantees (Mitchell et al. 2018) or did not scale (Ojha and Talukdar 2017). Recently, methods have emerged that estimate KG accuracy using small samples, providing strong statistical guarantees with minimal human efforts (Gao et al. 2019; Qi et al. 2022). Unlike them, we consider a multi-annotator scenario, developing strategies for routing triplets and aggregating labels based on annotators' expertise and reliability. Other works focus on validating useful triplets via crowdsourcing under a budget, but do not provide unbiased accuracy estimates with guarantees (Nguyen et al. 2019; Zheng et al. 2022). Our approach parallels these crowdsourcing practices by employing non-monetary rewards and using expert honey pots to assess layman annotators' quality.

DBpedia analyses. Previous efforts have been made to assess the quality of DBpedia. Acosta et al. (2013, 2018) used crowdsourcing to detect specific quality issues in DBpedia, analyzing common errors and classifying them for suitability in crowdsourcing tasks. Färber et al. (2018) established data quality criteria for evaluating and comparing several KGs, including DBpedia. However, these studies did not employ sampling strategies that could provide a representative evaluation of the entire KG, nor did they include CIs to ensure statistical guarantees. To the best of our knowledge, our work represents the most extensive effort in evaluating DBpedia accuracy, using a rigorous statistical technique.

LLM based assessment. Promising results have been achieved using LLMs to generate relevance judgments for information retrieval (Faggioli et al. 2023; MacAvaney and Soldaini 2023; Thomas et al. 2023). However, LLMs still fall short for KG accuracy evaluation (Mruthyunjaya et al. 2023; Sun et al. 2023). Auditing KG accuracy with LLMs differs fundamentally from generating relevance judgments, as verifying a single triplet correctness may require cross-referencing multiple (web) sources. This is more complex for an LLM due to potential noise and inaccuracies in the retrieved data. Our analyses further show LLMs limitations in auditing KG accuracy, highlighting their inability to retrieve learned knowledge when assessing triplet correctness.

Conclusions

In this work, we designed and implemented an efficient, utility-oriented KG accuracy evaluation framework capable of scaling to real-life KGs with limited human annotations and strong statistical guarantees. We applied this framework to DBpedia, introducing a utility-based strategy to partition the KG. By recruiting 60 layman annotators and three experts, we conducted the first extensive human-based accuracy evaluation of DBpedia, demonstrating the feasibility of a utility-based assessment procedure over a large KG.

Additionally, we explored the potential of LLMs in automating KG accuracy evaluation at scale. While the LLMs performed comparably to low-quality human annotators, they tended to overestimate accuracy scores, thereby not representing yet suitable replacements for human crowdworkers in auditing KG accuracy. As future work, we plan to further investigate the feasibility of making human annotators and LLMs collaborate to audit KG accuracy.

Acknowledgements

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074.

References

Acosta, M.; Zaveri, A.; Simperl, E.; Kontokostas, D.; Auer, S.; and Lehmann, J. 2013. Crowdsourcing Linked Data Quality Assessment. In *Proc. of ISWC 2013*, volume 8219 of *LNCS*, 260–276. Springer.

Acosta, M.; Zaveri, A.; Simperl, E.; Kontokostas, D.; Flöck, F.; and Lehmann, J. 2018. Detecting Linked Data quality issues via crowdsourcing: A DBpedia study. *Semantic Web*, 9(3): 303–335.

Alonso, O. 2019. *The Practice of Crowdsourcing*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers. ISBN 978-3-031-01190-0.

Arabzadeh, N.; Bigdeli, A.; and Bagheri, E. 2024. LaQuE: Enabling Entity Search at Scale. In *Proc. of ECIR 2024*, volume 14609, 270–285. Springer.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. G. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of ISWC 2007 + ASWC 2007*, volume 4825 of *LNCS*, 722–735. Springer.

Balog, K. 2018. *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*. Springer. ISBN 978-3-319-93933-9.

Batini, C.; Cappiello, C.; Francalanci, C.; and Maurino, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.*, 41(3): 16:1–16:52.

Bonifati, A.; Fletcher, G. H. L.; Voigt, H.; and Yakovets, N. 2018. *Querying Graphs*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Casella, G.; and Berger, R. L. 2002. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Cochran, W. G. 1977. *Sampling Techniques, 3rd Edition*. John Wiley. ISBN 0-471-16240-X.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20(1): 37–46.

Dalenius, T.; and Hodges, J. L. 1959. Minimum Variance Stratification. *J. Am. Stat. Assoc.*, 54(285): 88–101.

Dosso, D.; Davidson, S. B.; and Silvello, G. 2022. Credit distribution in relational scientific databases. *Inf. Syst.*, 109: 102060.

Esteves, D.; Rula, A.; Reddy, A. J.; and Lehmann, J. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *ACM J. Data Inf. Qual.*, 9(3): 16:1–16:26.

Faggioli, G.; Dietz, L.; Clarke, C. L. A.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; and Wachsmuth, H. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proc. of ICTIR 2023*, 39–50. ACM.

Faralli, S.; Lenzi, A.; and Velardi, P. 2023. A Benchmark Study on Knowledge Graphs Enrichment and Pruning Methods in the Presence of Noisy Relationships. *J. Artif. Intell. Res.*, 78: 37–68.

Färber, M.; Bartscherer, F.; Menne, C.; and Rettinger, A. 2018. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1): 77–129.

Gao, J.; Li, X.; Xu, Y. E.; Sisman, B.; Dong, X. L.; and Yang, J. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.*, 12(11): 1679–1691.

Garigliotti, D.; Albakour, D.; Martinez, M.; and Balog, K. 2019. Unsupervised Context Retrieval for Long-tail Entities. In *Proc. of ICTIR 2019*, 225–228. ACM.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Hasibi, F.; Balog, K.; and Bratsberg, S. E. 2017. Dynamic Factual Summaries for Entity Cards. In *Proc. of SIGIR 2017*, 773–782. ACM.

Hasibi, F.; Nikolaev, F.; Xiong, C.; Balog, K.; Bratsberg, S. E.; Kotov, A.; and Callan, J. 2017. DBpedia-Entity v2: A Test Collection for Entity Search. In *Proc. of SIGIR 2017*, 1265–1268. ACM.

Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.*, 194: 28–61.

Hung, N. Q. V.; Nguyen, T. T.; Lam, N. T.; and Aberer, K. 2013. An Evaluation of Aggregation Techniques in Crowdsourcing. In *Proc. of WISE 2013*, volume 8181 of *LNCS*, 1–15. Springer.

Ilievski, F.; Vossen, P.; and Schlobach, S. 2018. Systematic Study of Long Tail Phenomena in Entity Linking. In *Proc. of COLING 2018*, 664–674. Association for Computational Linguistics.

Ilyas, I. F.; Lacerda, J.; Li, Y.; Minhas, U. F.; Mousavi, A.; Pound, J.; Rekatsinas, T.; and Sumanth, C. 2023. Growing and Serving Large Open-domain Knowledge Graphs. In *Proc. of SIGMOD 2023*, 253–259. ACM.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M. A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proc. of ACL 2022*, 3214–3252. ACL.

MacAvaney, S.; and Soldaini, L. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *Proc. of SIGIR 2023*, 2230–2235. ACM.

Marchant, N. G.; and Rubinstein, B. I. P. 2017. In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling. *Proc. VLDB Endow.*, 10(11): 1322–1333.

Marchant, N. G.; and Rubinstein, B. I. P. 2021. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. In *Proc. of KDD 2021*, 1180–1190. ACM.

Marchesin, S.; and Silvello, G. 2024. Efficient and Reliable Estimation of Knowledge Graph Accuracy. *Proc. VLDB Endow.*, 17(9): 2392–2404.

Marchesin, S.; Silvello, G.; and Alonso, O. 2024. Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. In *Proc. of CIKM 2024*. ACM.

Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G. C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J. B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; and Chiu, J. 2024. Gemma: Open Models Based on Gemini Research and Technology. *CoRR*, abs/2403.08295.

Mitchell, T. M.; Cohen, W. W.; Jr., E. R. H.; Talukdar, P. P.; Yang, B.; Betteridge, J.; Carlson, A.; Mishra, B. D.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E. A.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R. C.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2018. Never-ending learning. *Commun. ACM*, 61(5): 103–115.

Mohoney, J.; Pacaci, A.; Chowdhury, S. R.; Mousavi, A.; Ilyas, I. F.; Minhas, U. F.; Pound, J.; and Rekatsinas, T. 2023. High-Throughput Vector Similarity Search in Knowledge Graphs. *Proc. ACM Manag. Data*, 1(2): 197:1–197:25.

Mruthyunjaya, V.; Pezeshkpour, P.; Hruschka, E.; and Bhutani, N. 2023. Rethinking Language Models as Symbolic Knowledge Graphs. *CoRR*, abs/2308.13676.

Nguyen, T. T.; Yin, H.; Weidlich, M.; Zheng, B.; Nguyen, Q. V. H.; and Stantic, B. 2019. User Guidance for Efficient Fact Checking. *Proc. VLDB Endow.*, 12(8): 850–863.

Ojha, P.; and Talukdar, P. P. 2017. KGEval: Accuracy Estimation of Automatically Constructed Knowledge Graphs. In *Proc. of EMNLP 2017*, 1741–1750. ACL.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *CoRR*, abs/2306.08302.

Paranjpe, A. P.; Bhowmik, R.; and de Melo, G. 2020. Facts That Matter: Dynamic Fact Retrieval for Entity-Centric Search Queries. In *Proc. of ISWC 2020*, volume 2721 of *CEUR Workshop Proceedings*, 310–315. CEUR-WS.org.

Paulheim, H. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, 8(3): 489–508.

Qi, Y.; Zheng, W.; Hong, L.; and Zou, L. 2022. Evaluating Knowledge Graph Accuracy Powered by Optimized Human-Machine Collaboration. In *Proc. of SIGKDD 2022*, 1368–1378. ACM.

Reinanda, R.; Meij, E.; and de Rijke, M. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Found. Trends Inf. Retr.*, 14(4): 289–444.

Samadi, M.; Talukdar, P. P.; Veloso, M. M.; and Mitchell, T. M. 2015. AskWorld: Budget-Sensitive Query Evaluation for Knowledge-on-Demand. In *Proc. of IJCAI 2015*, 837–843. AAAI Press.

Settles, B. 2009. Active Learning Literature Survey.

Stadler, C.; Saleem, M.; Mehmood, Q.; Buil-Aranda, C.; Dumontier, M.; Hogan, A.; and Ngomo, A. C. N. 2024. LSQ 2.0: A linked dataset of SPARQL query logs. *Semantic Web*, 15(1): 167–189.

Suchanek, F.; Alam, M.; Bonald, T.; Chen, L.; Paris, P.-H.; and Soria, J. 2024. YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. In *Proc. of SIGIR 2024*. ACM.

Sun, K.; Xu, Y. E.; Zha, H.; Liu, Y.; and Dong, X. L. 2023. Head-to-Tail: How Knowledgeable are Large Language Models (LLM)? A.K.A. Will LLMs Replace Knowledge Graphs? *CoRR*, abs/2308.10168.

Thomas, P.; Spielman, S.; Craswell, N.; and Mitra, B. 2023. Large Language Models can Accurately Predict Searcher Preferences. *CoRR*, abs/2309.10621.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M. A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Vrandecic, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10): 78–85.

Xue, B.; and Zou, L. 2023. Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.*, 35(5): 4969–4988.

Zhang, J.; Wu, X.; and Sheng, V. S. 2016. Learning from crowdsourced labeled data: a survey. *Artif. Intell. Rev.*, 46(4): 543–576.

Zheng, L.; Cheng, P.; Chen, L.; Yu, J.; Lin, X.; and Yin, J. 2022. Crowdsourced Fact Validation for Knowledge Bases. In *Proc. of ICDE 2022*, 938–950. IEEE.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-Following Evaluation for Large Language Models. *CoRR*, abs/2311.07911.